International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.7, No.4, April 2018 DOI:10.15693/ijaist/2018.v7i4.1-7

A modified ordered K-means clustering based Scholar's Performance Forecast Analysis

P. Sundari[1], Assistant Professor
[2]N.Muneeslakshmi, M.Phil Research Scholar,
[1,2]Department of Computer Science
Government Arts College (Autonomous) Coimbatore-18, Tamil Nadu

Abstract

Development of an instructive establishment can be measured with respect to the effective scholars of the organization. Hence, this investigation identifies with the expectation of student execution in higher instruction medium that appears as an elementary prerequisite for making change in quality instruction. The data mining procedures plays an essential part in information investigation. For the development of the organization, K-means based algorithm related with the mining have been utilized here as a part of the exploration. Various external factors are there to affect the execution of studies. The main utilization of the K-means is to cluster each attributes effectively related to the students' performance. In this paper, modified ordered K-means clustering algorithm is intended to predict the best reasonable outcome to demonstrate the development. Results got in the present investigation might be supportive for distinguishing the attributes that could help to take some activities and achievement rate could be expanded adequately. The overall results will be carried out by MATLAB simulation tool which represents that the investigation might be supportive for distinguishing the learners that could take some activities and achievement rate could be expanded adequately. It is finally summarized as the proposed ordered K-means based clustering helps to find the exact value similar to the original value mentioned in dataset.

Keywords--- Classification, Data mining, modified ordered K-means clustering algorithm, Prediction, Cross validation.

I. Introduction

Data mining always helps to quote the relevant information from the large datasets [1]. Its practises are always helpful for data analysis and predictions. Particularly, the classification is an unsupervised learning practise that helps to classify the predefined class labels. There is an "N" number of numerous traditional classification techniques such as Decision tree algorithm, Bayesian network, neural network and Genetic algorithm etc. Hence, the traditional methods are utilized to build the classification model for predicting the future trend based on previous pattern.

Clustering has been used in a number of applications such as engineering, biology, medicine and data mining. The most popular clustering algorithm used in several field is K-Means since it is very simple, fast and efficient. K-means is developed by Mac Queen. The K-Means algorithm is effective in producing cluster for many practical applications. But the computational complexity of the original K-Means algorithm is very high, especially for large datasets. The K-Means algorithm is a partition clustering method that separates data into K groups. Main drawback of this algorithm is that of a priori fixation of number of clusters and seeds.

To rectify the drawbacks of K-means algorithm a new algorithm is proposed namely modified Ordered K-means Clustering which starts its computation without representing the number of clusters and the initial seeds. The proposed clustering algorithm purely works on affinity measure which helps to fix the number of resultant clusters. It divides the dataset into some number of clusters with the help of threshold value. The uniqueness of the cluster is based on the threshold value. The number of clusters increases on decreasing the threshold value and the number of cluster decreases by increasing the threshold value. More unique cluster is obtained when the threshold value is smaller.

Nowadays, in many institutions the maintenance of the large database is complex to maintain several attributes such as grade, character, fitness, sports, etc. Apart from these aspects, the extra circular activities are also added with academic database. Hence, its complexity will be increasing day by day because of huge number of attributes. Therefore, data base management is considered in managing the data to discover such information and knowledge about students. The analysis discovered for maintaining and extracting the knowledge under different techniques namely, machine learning, visualization and statistical techniques. The simple research methodology is shown in the figure 1.

The main contribution of this research is to analyse the K-means algorithm and modify the internal features corresponding to the dataset attributes. The designed algorithm follows on different strategy with simple steps. The attributes considered here are age, travel time, study time and health. The algorithm is modified with simple steps as represented. The process helps to identify the exact numerical attributes.

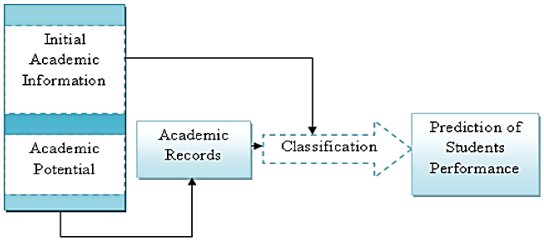

Figure 1: Block diagram of classification

The remaining part of this paper is described as follows. Section 2 describes the recent data prediction

analysis utilized in different aspects. In section 3, the proposed research methodology is described with neat flow diagram. To verify the prediction analysis, the experimental results carried out by MATLAB tool is described under section 4. Finally, the paper is summarized in section 5.

II. Literature survey

Romero et al., (2013) presented the web usage mining in e-learning systems to predict the final exam marks that the university students will obtain. It also constructed a specific Moodle mining tool for newcomers like instructors and courseware authors. The execution of various information digging systems for grouping students are thought about, beginning with the understudy's use information in a few Cordoba University Moodle courses in engineering. A few understood characterization strategies have been utilized, for example, statistical methods, decision trees, rule and fuzzy rule induction methods, and neural networks. It also carried out a research over available and filtered data to try to obtain more accuracy. At last, some models are considered and described with a classifier model appropriate for an excellent educational environment for accurate and comprehensible order for instructors to handle excellent decision making.

Hostetler (1983) examines to what extent a student's aptitude in computer programming may be predicted through measuring certain cognitive skills, personality traits and past academic achievement. The customary classroom conditions are the most broadly utilized instructive structures. It depends on eye to eye contact amongst instructors and students and composed in addresses. There are various subtypes: private and state funded instruction, basic and essential training, grown-up training, higher, tertiary and scholarly instruction, custom curriculum, and so on. In regular classrooms, instructors endeavor to improve guidelines by checking understudy's learning forms and breaking down their exhibitions utilizing paper records and perception. They can likewise utilize data about understudy participation, course data, educational programs objectives, and individualized arrangement information. In conventional instructive situations, order has been connected for some errands, a few cases are: anticipating understudy achievement utilizing various relapse conditions in an early on programming course, bearing in mind the end goal of better guiding for students.

Dietz-Uhler and Hurn (2013) defined the learning analytics in educational institutions. It states how to predict the learning process to monitor and predict the student performance. It also discusses several issues and concerns with the utilization of learning analytics in higher education. Likewise, it is important to manage the classification process under different strategies.

Ahmed and Elaraby (2014) proposed the classification task is used to predict the final grade of students and as there are many approaches that are used for data classification, the decision tree (ID3) method is used here. Bhargava et al., (2013) handled decision tree analysis on j48 algorithm. The predicting tasks are identified with the help of Fisher and Schlimmer. Recent publishers discussed about the state of art about practical machine learning tools and techniques, witten et al., (2016).

Coussement et al., (2017) presented the data preparation on customer churn prediction performance. With this motivation several other reviews are analysed such as Márquez-Vera et al., (2016), Ekanadham and Karklin (2017), Ameri et al., (2016) and Ognjanovic et al., (2016).

D.Napoleon, S.Pavalakodi (2011) proposed dimension reduction concept in K-Means clustering algorithm. Clustering is the process of finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups [15].

Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality that corresponds to the intrinsic dimensionality of the data. K-Means clustering algorithm often does not work well for high dimension, hence, to improve the efficiency, apply Principal components analysis (PCA) on original data set and obtain a reduced dataset containing possibly uncorrelated variables. In this paper, principal component analysis and linear transformation is used for dimensionality reduction and initial centroid is computed, then it is applied to K-Means clustering algorithm.

III. Research methodology

Initially, the research is identified with the k-means clustering concept and based on the input attributes, the process is varied with different computation factors. In this section 3.1, the K-means clustering is identified and described. The future extension of this algorithm with ordered clustering is described in the section 3.

3.1 K-Mean Clustering

The main objective of cluster is to group the object that are similar in one cluster and separate objects that are dissimilar by assigning them to different clusters. One of the most popular clustering methods is K-Means clusters algorithm. It classifies objects to pre-defined number of clusters, which is given by the user (assume K clusters). The idea is to choose random cluster centers, one for each cluster. These centers are preferred to be as far as possible from each other. In this algorithm Euclidean distance measure is used between two multidimensional data points.

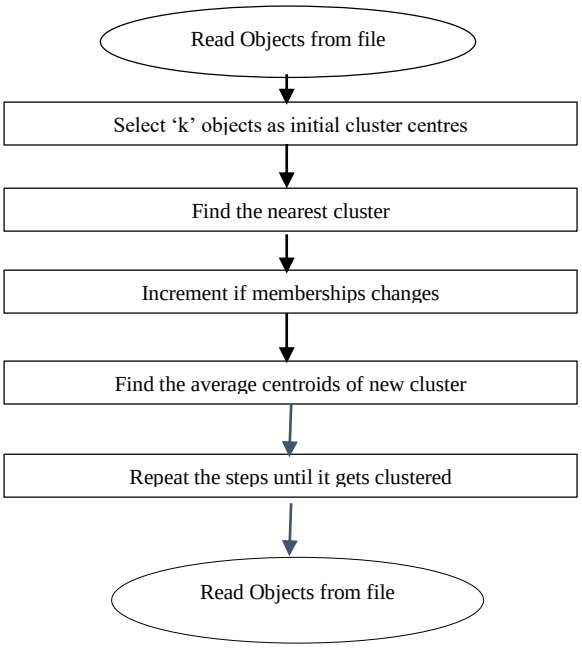

Figure 2: Traditional K-means clustering algorithm

The K-Means method aims to minimize the sum of squared distances between all points and the cluster centre. The algorithmic steps are described in the following

The pseudo code for the K-means clustering is defined as follows:

Initialize the data objects, clusters, memberships
Find the threshold value and find the cluster head
While (N> threshold)
For 1 tends to 0 to N-1
If distance <dmin
Perform membership function
Find new cluster head and process
Select the cluster size and display

The major limitations are noticed in K-means is listed as follows:

- It is very difficult to predict the number of clusters with initial seeds have a strong impact on the final results.
- In some cases, the numerical values are approximated to its nearest value. It leads in severe variation in the graphical results.
- It is not suitable for large datasets because the identification of cluster head is difficult to identify.

3.2. Proposed Ordered Clustering

The ordered clustering is utilized in this student’s database to classify the result with high accuracy. In this research, a special clustering problem called ordered clustering problem was utilized which is addressed by Smet and Gilbart [28] for dealing with the country risk evaluation problem. As previously noted, the identification of the ordered clusters can provide a necessary support for the DM to sort the alternatives. Unlike the classical clustering problem, the ordered clustering problem not only partitions the alternatives into the predetermined number of clusters, but also has a completely ranking relationship of these clusters.

Figure 3 provides the flow chart for proposed ordered K-means clustering algorithm. Moreover, the current K-mean clustering calculations are essentially used to group the information into a few gatherings which don't have any connection among them. In multi-criteria decision aid (MCDA), the leader may want to get "requested bunches" in which there exist the requested relations among the groups. This sort of issues can be alluded to as multi-criteria requested grouping issues.

The recognizable proof of requested groups can give the need relations of choices for the DMs. In spite of the fact that the requested bunches can't give more precise relations of options than the entire rankings of the considerable number of options, the requested grouping is additionally fundamental in the genuine issues. For instance, in the positioning of world colleges, the DMs may not give the exact rankings of a few colleges since they have no conspicuous contrasts. In this way, it is sensible that the options with no huge contrasts into the requested bunches.

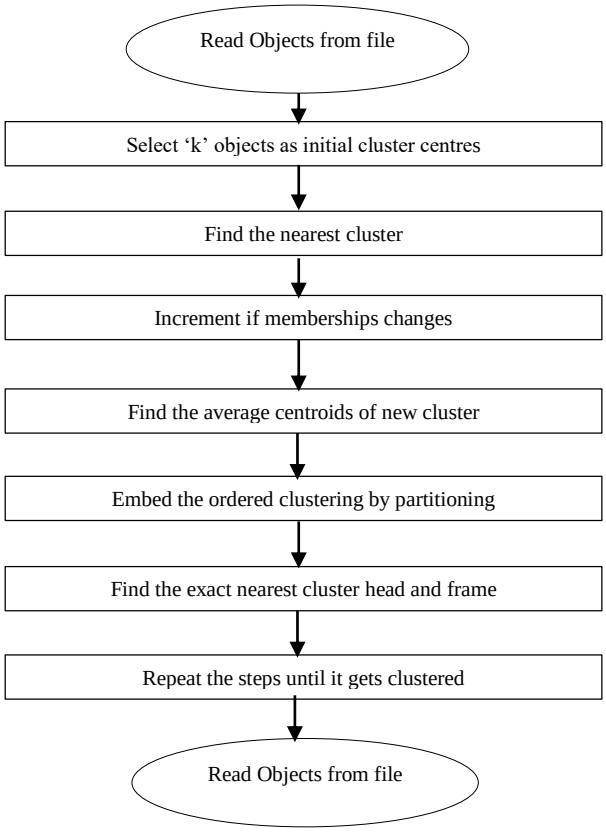

Figure 3: Proposed modified ordered K-means clustering

From the figure 3, the k-means procedure is added with the partitioning model by combining the complex clustering algorithm to measure the exact cluster head. It provides exact membership functions to measure the attributes. The difference between the

traditional and the proposed clustering is that partitioning technique. It cluster large number of datasets with respect to the input attributes. The pseudo code for the proposed ordered clustering implemented in clustering the complex dataset.

Pseudo code for K-means clustering algorithm

Initialize the process of attributes (f,p,ds)
Select 'k' data objects representing the cluster centroids
Get the variables like 'age', 'travel time', 'study time', 'free time', 'health'
Read all the attributes
Assign each data object of the entire data set to the cluster having the closest centroid
while hasdata(ds)
T = read(ds);
End
Compute new centroid for each cluster
Initiate hybrid k-means clustering (Partioning based method)
Hierarchial divisive approach ordered clustering
plotting data
while hasdata(ds)
T = read(ds);
End
If one of the centroids gets mismatched then check until it gets matched
Display the classification process

In the proposed algorithm, the exact original age is identified with respect to the cluster head and centroid. A prototype module of an exact data set has been estimated. It is noticed that the implementation simplifies the numerical values of the data set with their coordinates between 0 and 1. The traditional K-means algorithm depends on the numerical access with approximation. It provides the real time error and merge previous data to present analysis. Hence, the overlapping is occurred.

To prevent such issues, a hybrid model is attached to this K-means and ordered in a flow based on the priority of the nearby values. It is completely based on the uniform distribution generator. It process 4, 8 and 12 centroids with the standard K-Means algorithm and finally reduces the analysis from 10% to 13%. Hence, it is confirm that the improvement shown by this ordered model prototype helps to predict the values in complex dataset.

a) Dataset

Cortez and Silva (2008) framed the dataset and published it in Paulo Cortez, University of Minho, GuimarÃ£es, Portugal,at the link "http://www3.dsi.uminho.pt/pcortez". It contains 33 attributes, in that only 5 specific attributes are selected for processing. The dataset which is considered for processing this proposed algorithm is taken from the secondary school student performance. Here the major attributes are shown in the figure 4. It is readable file for processing the input csv file. The attributes are selected based on the Setting a target exam, grade points and some personnel details as shown in table 1. The attributes are separated with 0 and 1 statement. The evaluation model is completely carried out with the MATLAB simulations.

Figure 4: Dataset Considered for evaluation in Excel format (.csv format)

Table 1 List of attributes considered from dataset

Attribute Name	Attribute Type
Age	Numeric: from 15 to 22
Travel time	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
Study time	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
Free time	Free time after school (numeric: from 1 - very low to 5 - very high)
Health	current health status (numeric: from 1 - very bad to 5 - very good)

a) Experimental results

The performance analysis of the proposed method is analysed based on the input attributes. The numerical data is collected from the dataset with the age bar from 15 to 22. In traditional K-means clustering, the classification of age is separated as shown in the figure 5.

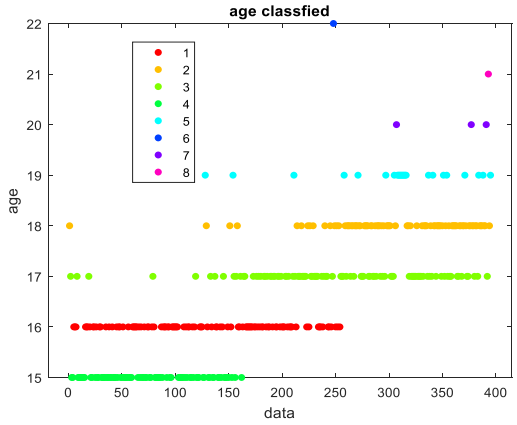

Figure 5: Representation of age classification based on K-means algorithm

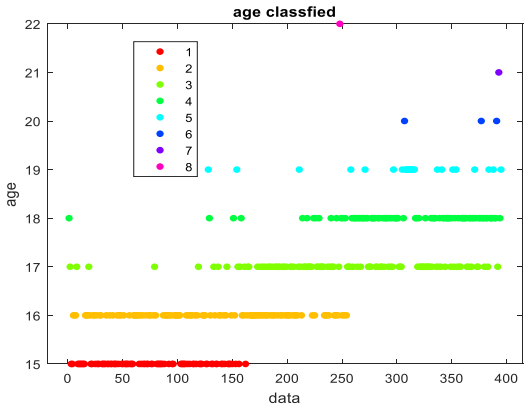

Figure 6: Representation of age classification based on proposed ordered clustering algorithm

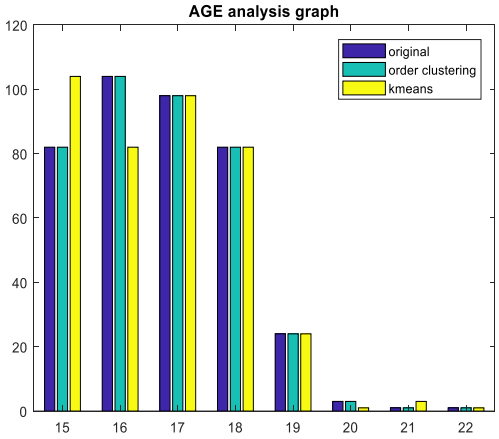

Figure 7: Comparison of accuracy with original data, K-means and Proposed Ordered clustering

Figure 7 segments the age regarding with the students. It provides the distribution of 395 students. With varies trails the evaluation is carried out with the bar graph representations. Hence, the classification accuracy of this cases is denoted as error free operation. It proves that the computational complexity is reduced even at large datasets. Similarly, in future if the amount of student's dataset increases then the classification accuracy will be improved. The result proves that the accuracy of the proposed ordered clustering concept is resulted in improved status.

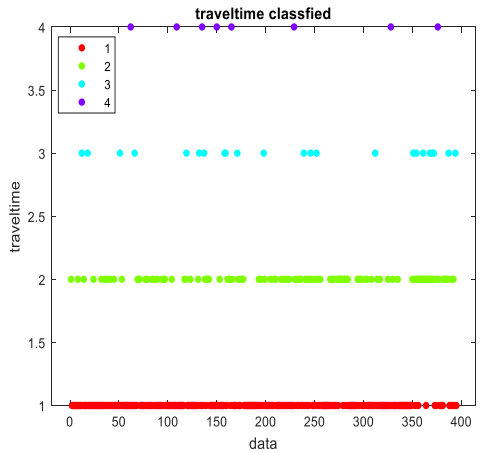

Figure 8: Travel time analysis for existing K-means

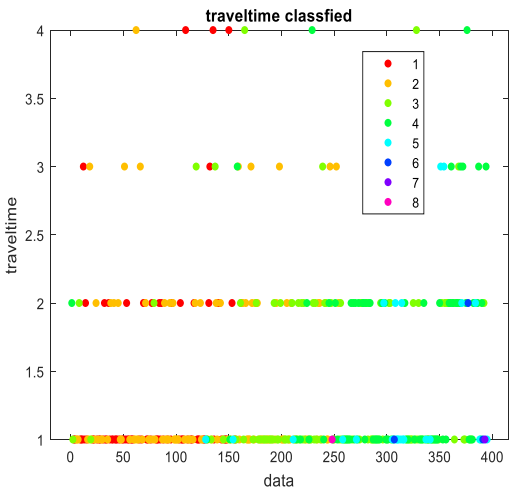

Figure 9: Travel time analysis for proposed ordered clustering

Similar to the previous step, the traveling time is identified from respective attribute values. The respective colours are indicating the traveling time classification with the help of data inputs. Figure 8, 9 and 10 displays the travel time of the proposed ordered clustering and the original data values are equal. It states that the proposed module has the prediction accuracy as high and error free operation.

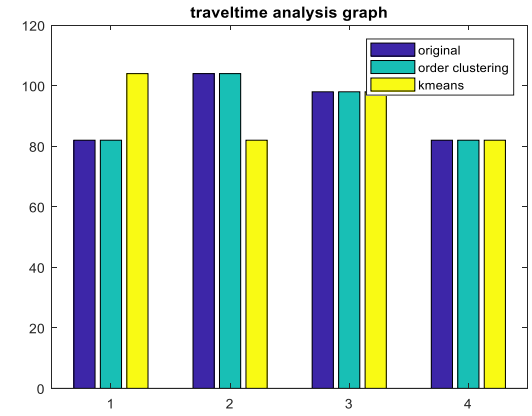

Figure 10: Travel time analysis of existing K-means and proposed ordered clustering

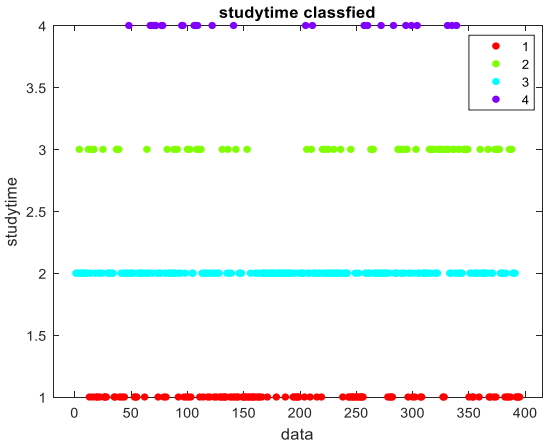

Figure 11 Study time analysis of K-means

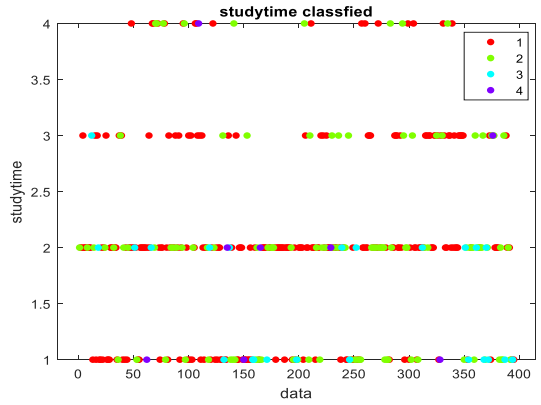

Figure 12: Study time analysis of Proposed ordered clustering

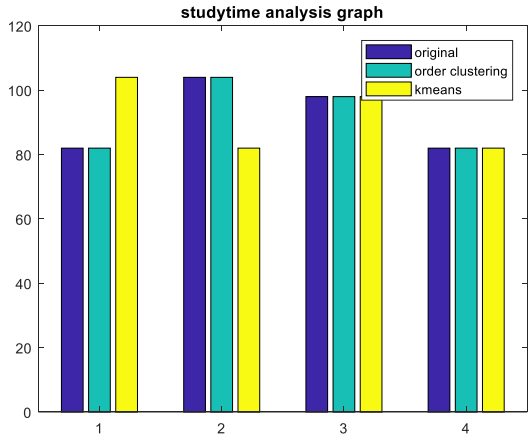

Figure 13: Proposed classification is compared with the original value and the existing value

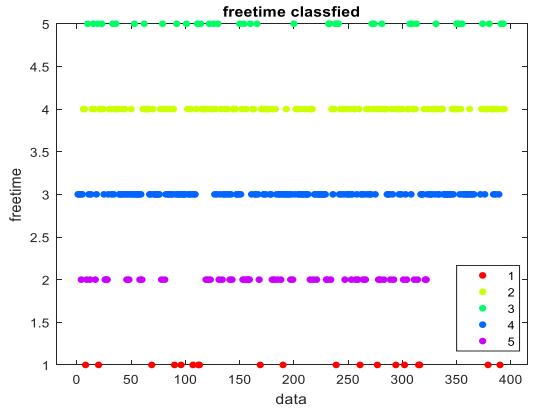

Figure 14: Free time Classification of existing K-means and proposed ordered clustering

It is easy to implement large number of variables and results in faster computation. The health analysis is one of the numeric concept that classifies the data from 1 to 5. It is noticed that the ordered clustering provides exact prediction related to the original reference data.

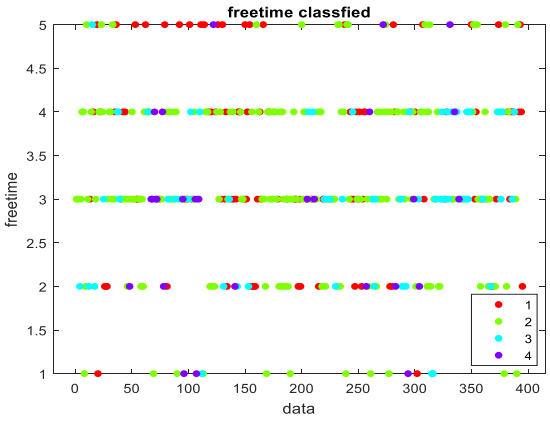

Figure 15: Free time Classification of Proposed ordered clustering

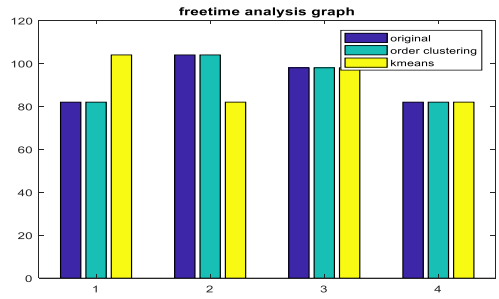

Figure 16: Accuracy for existing K-means and proposed Ordered Clustering free time analysis compared with the original value.

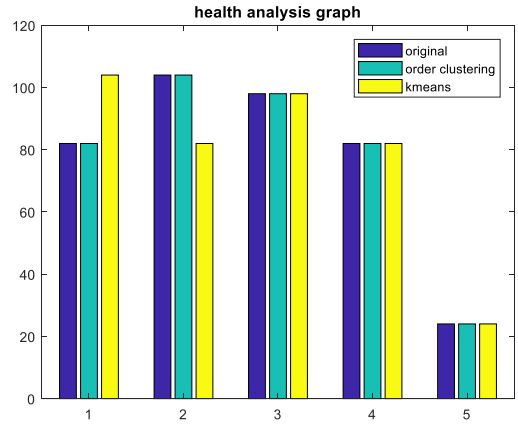

Figure 17: Accuracy for existing K-means and proposed Ordered Clustering health analysis compared with the original value.

From the experimental results it is proved that the proposed procedure demonstrated with the groups of learner's execution. The classifiers has been worked by consolidating the standard for Data Mining that incorporates learner's execution with the utilization of information mining strategies. In another aspect, utilizing this ordered clustering provides the accomplishment of good score. From the experimental results, the sensitivity is improved and it is suitable for large datasets. The time complexity is totally reduced. The exact prediction decides the classifier accuracy.

IV. CONCLUSION

The research work has been carried out to enhance the accuracy of data prediction by using K-means clustering. In web mining, the analysis of high dimensional data is difficult to classify the complex data. Hence, this work contributes the classification of are age, travel time, study time and health. The dataset is collected from an institution is with 395 students. The dataset samples are collected and processed with the MATLAB simulation tool. The accuracy of the proposed classified tests are increased when compared with the traditional k-means clustering method. Results proves that the proposed method is best suitable for complex dataset.

V. Future work

An interesting issue is that analysts have yet not demonstrated that when the labelled data set is adequately large, the unlabeled information still offer assistance. This can fill in as the helpful future direction. There is likewise a need to assess the convenience of unlabeled cases in different spaces and on various datasets keeping in mind the end goal to reach to some solid conclusions.

References

[1]. Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques, 2nd editionMorgan Kaufmann Publishers. *San Francisco, CA, USA*.

[2]. Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., & Ventura, S. (2013). Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education*, *21*(1), 135-146.

[3]. Hostetler, T. R. (1983). Predicting student success in an introductory programming course. *ACM SIGCSE Bulletin*, *15*(3), 40-43.

[4]. Dietz-Uhler, B., & Hurn, J. E. (2013). Using learning analytics to predict (and improve) student success: A faculty perspective. Journal of Interactive Online Learning, 12(1), 17-26.

[5]. Rokach, L., & Maimon, O. (2014). *Data mining with decision trees: theory and applications*. World scientific.

[6]. Ahmed, A. B. E. D., & Elaraby, I. S. (2014). Data Mining: A prediction for Student's Performance Using Classification Method. *World Journal of Computer Application and Technology*, *2*(2), 43-47.

[7]. Fisher, D. H., & Schlimmer, J. C. (2014, May). Concept simplification and prediction accuracy. In *Proceedings of the Fifth International Conference on Machine Learning* (pp. 22-28).

[8]. Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(6).

[9]. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[10]. Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, *95*, 27-36.

[11]. Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, *33*(1), 107-124.

[12]. Ekanadham, C., & Karklin, Y. (2017). T-skirt: Online estimation of student proficiency in an adaptive learning system. *arXiv preprint arXiv:1702.04282*.

[13]. Ameri, S., Fard, M. J., Chinnam, R. B., & Reddy, C. K. (2016, October). Survival analysis based framework for early prediction of student dropouts. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 903-912). ACM.

[14]. Ognjanovic, I., Gasevic, D., & Dawson, S. (2016). Using institutional data to predict student course selections in higher education. *The Internet and Higher Education*, *29*, 49-62.

[15]. Napoleon, D., & Pavalakodi, S. (2011). A new method for dimensionality reduction using k-means clustering algorithm for high dimensional data set. *International Journal of Computer Applications*, *13*(7), 41-46.

Author Details

P. Sundari received the B.E. degree in Computer Science Engineering in 1992, from Government College of Technology College, Coimbatore India. She received M.C.A degree in 1998 from Bharathiar University, Coimbatore, India and completed MPhil degree in 2009 from Kongunadu Arts& Science College in Coimbatore. Currently, she is working as an Assistant Professor in Department of Computer Science at Government Arts College (Autonomous) Coimbatore. India.

N. Munees Lakshmi received the B.C.A degree in Computer Science Engineering in 2014 from Bharathiar University, valparai, India and M.SC. Degree in Information Technology in 2016 from Michael job Arts& Science College, Sulur, India. Currently she pursuing as an Mphil Research Scholar in Government Arts College (Autonomous) Coimbatore, India.